

Original papers

Zero shot multistage hierarchical fish classification using YOLOE and large language model

Jisha Anu Jose ^a, Renjith Thomas ^b^a Quest Global Engineering Services Pvt. Ltd., India^b St. Joseph's College of Engineering and Technology, Palai, India

ARTICLE INFO

Keywords:

Object detection
Zero shot prompt
Multimodal large language models
Classification
Evaluation

ABSTRACT

Manual classification of fish species is challenging, time-consuming, and requires experience, especially when dealing with similar species. To train models for the development of an automated categorization system, large datasets are required. A zero-shot fish species classification system is proposed to categorize fish species with limited data availability. The system combines a multi-mode large language model with an object detection model. A visual-prompted YOLOE object detection model is used to detect fish from the images in their natural environment. By identifying the fish in the image, the model enables separation of the fish from the background. The fish region is forwarded to a multi-stage hierarchical classification phase after being cropped from the background. For the categorization, a large language model with zero-shot prompts is employed. A physical description of the fish species is included in the prompt for each classification level. The classification stages are evaluated using metrics such as accuracy, precision, recall, and F1-score; mean average precision is utilized as an additional statistic for object detection. YOLOE's zero-shot object detection accuracy was found to be 88%, while the large language model's multi-level species categorization accuracy was 87.20%. The proposed zero-shot pipeline is also compared with the state-of-the-art-systems and found that it performed better than the CLIP zero-shot baseline and nearly reached the performance of fully supervised models.

1. Introduction

Fish vary greatly in size, shape, and color; some species can grow to several meters in length as adults. They are available in a variety of body shapes, such as fusiform, elongated, compressed, and depressed morphologies. Determining species abundance in certain habitats requires precise fish species identification (Malla et al., 2022). A country's GDP is greatly impacted by the export of commercial and ornamental fish as well as other fisheries products. In order to evaluate stocks and export processed fish and fish products, commercial and ornamental fish must be classified (Ghaly et al., 2013). Traditionally, professionals differentiate between species using morphological traits they have learned through close study. Fish sorting is still a time-consuming and labor-intensive process, though, both on land and on fishing boats. Morphometric techniques for species identification have been improved by advances in image processing (Cadriin and Friedland, 1999). Fish species can be quickly and effectively separated using automated image-based technology, increasing operational effectiveness and lowering spoilage.

To minimize human labor and facilitate quicker decision-making, computer vision has been extensively utilized for automated fish species

categorization (Tejaswini et al., 2024; Jose et al., 2022; Rachmatullah and Supriana, 2018). The majority of early methods depended on conventional machine learning models with features extracted using different feature descriptors. For example, Tamou et al. (2018) proposed a linear SVM-based underwater fish categorization system. A genus-species classification method using a customized dataset reported 98.03% and 93.14% accuracy at the genus and species levels, respectively, with ELM-based kernels (Jose and Kumar, 2020). While OFDNet, based on the SSD architecture, achieved 89.7% accuracy for three species in foggy underwater conditions (Christensen et al., 2018).

Deep learning has proven to be quite effective in extracting hierarchical representations from visual data (Guo et al., 2016). Several Convolutional Neural Network (CNN)-based studies have demonstrated excellent performance in fish categorization. Even with insufficient data, a VGG16-based model trained on blended RGB pictures obtained a true acceptance rate of 96.4% on the QUT dataset (Hridayami et al., 2019). This CNN model also attained an accuracy of 99.7% on the improved LifeCLEF15 dataset. Inception-v3, ResNet50, VGG19, and fine-tuned VGG16 were compared in dos Santos and Gonçalves (2019), with Inception-v3 attaining 0.96, 0.938, and 0.873 accuracy at order,

* Corresponding author.

E-mail address: jishaanujose@gmail.com (J.A. Jose).<https://doi.org/10.1016/j.compag.2025.111288>

Received 27 May 2025; Received in revised form 17 November 2025; Accepted 1 December 2025

Available online 16 December 2025

0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

family, and species levels, respectively. Several studies have investigated integrated detection and classification processes in addition to classification. Using the NCFM dataset, a You Only Look Once (YOLO) v2-based real-time system classified eight species onboard with a mAP of 0.912 (Wang et al., 2018). While Kandimalla et al. (2022) used YOLOv3/YOLOv4 with Norfair-based tracking for eight species, obtaining a mAP of 0.73, Knausgård et al. (2022) suggested a two-stage pipeline utilizing YOLOv3 for detection followed by a SE-Net-based classifier. For fish categorization, transformer-based models have also been investigated. Fish-TViT demonstrated the importance of token-based attention by achieving 94.33% and 98.34% accuracy on Fishnet and Fish-Pak datasets with 12 and 6 species, respectively (Gong et al., 2023), while token selective vision transformer achieved 74.3%, 90.4%, and 94.7% accuracy for plankton, shark, and wildfish datasets (Si et al., 2023).

The application of Large Language Models (LLMs) in computer vision tasks has recently extended beyond text comprehension. For example, GPT-4o was evaluated for image classification using the Fashion-MNIST dataset (Wu et al., 2025). Yang et al. (2024) introduced a data-free framework for multi-label image recognition, employing textual prompt engineering to adapt CLIP for multi-label tasks. Pratt et al. (2023) demonstrated the effectiveness of LLM-assisted visual classification by assessing zero-shot open-vocabulary image classification combined with LLM reasoning across 15 benchmark datasets. Additionally, FishDetectLLM (Zhu et al., 2025) explored the use of Tiny-LLaVA to simultaneously perform fish detection and species identification by leveraging the multimodal reasoning capabilities of LLMs.

According to recent literature, fish classification methods have progressed from traditional machine learning to CNN-based deep learning, transformer architectures, and, most recently, LLM-based vision-language models. However, most current systems achieve high accuracy by relying on large, annotated training datasets, controlled imaging conditions, or species-specific retraining. Consequently, these approaches are less suitable for under represented species with limited annotated data. Few studies have investigated zero-shot multimodal LLM-based strategies tailored for underwater organisms in natural environments. Zero-shot systems (Song et al., 2023) offer a promising solution to this limitation, as they do not require dataset-specific training.

Addressing these limitations, this study presents a zero-shot system that integrates YOLO-based open-vocabulary fish identification with a large language model driven multi-stage classification approach. The method operates without the need for training data, utilizing both visual and textual prompts to identify fish in natural underwater imagery and classify species. This flexibility is particularly advantageous in scenarios involving underrepresented species. YOLO-based detectors are selected for their robust performance under challenging conditions such as low lighting, partial occlusion, and cluttered backgrounds. Furthermore, the zero-shot reasoning capabilities of contemporary LLMs address the scarcity of underwater fish images and enable scalable identification of uncommon species without dataset-specific training.

The main objectives of this study are as follows:

- To create a zero-shot hierarchical fish identification pipeline that combines open-vocabulary YOLO-based fish detection and LLM-based species classification with stage-wise prompts.
- To compare the zero-shot LLM-based framework with traditional supervised baselines (ResNet50, EfficientNet-B0) and the zero-shot ViT-L/14, particularly in low-data circumstances.

The proposed system aims to provide a scalable, annotation-free alternative to existing supervised fish classification methods that can be deployed in real-world marine biodiversity monitoring systems.

Section 2 describes the proposed system and provides a detailed description of each step. Section 3 discusses and compares the system's performance. The work is concluded in Section 4.

2. Proposed methodology

2.1. Outline of the system

The differentiation of fish species is essential for effective stock assessment and ecological research. Observing fish within their natural habitats enables more accurate stock evaluations. This study focuses on coral reef fish, as their population levels are frequently used as indicators of reef ecosystem health. Therefore, precise identification of coral reef fish species is vital for both population assessments and export evaluations.

Coral fish are brightly colored fish that closely resemble their native habitats. It is challenging to identify these fish in their natural habitat. The suggested approach is based on images of fish taken in their natural habitat. The dataset includes images of eight different fish species. These images are analyzed using a visual and text-prompted YOLOE model (Wang et al., 2025), which detects and localizes fish locations. The spotted fish regions are then cropped from the background and proceed to the next classification stage.

A multi-stage classification method is implemented for separation of species. Initially, cropped images are sorted into five categories, with four visually similar species combined into a "Stripped class" category. Those images classified under the "Stripped class" proceed to the second stage, where they are further divided into three sub-categories, with two species grouped into a "White striped" class. In the final stage, the "White striped" class is broken down into individual species. Each classification level employs a specific zero-shot prompt detailing the descriptive attributes of the classes at that stage. The reasoning behind creating these intermediate grouped classes during the classification process is thoroughly explained in the manuscript. The system's schematic representation is shown in Fig. 1.

2.2. Dataset details

The FishBase dataset (Froese et al., 2010), an extensive global information system on fin fishes, was used for system evaluation. From the over 33,000 species cataloged, eight were selected for the investigation, including *Abudefduf Vaigiensis*, *Acanthurus Nigrofusculus*, *Amphiprion Clarkii*, *Chaetodon Speculum*, *Chaetodon Trifascialis*, *Dascyllus Aruanus*, *Hemigymnus Melapterus*, and *Myripristis Kuntze*. Some images in the study depict multiple fish of the same species. Details about the dataset, such as the number of images and sample visuals, are presented in Table 1. Since YOLOE and the LLM were used in zero-shot mode, there was no training or fine-tuning of the models, which made the usual training-validation-test data splits unnecessary.

2.3. Zero shot object detection using YOLOE model

A novel zero-shot promptable YOLO model named YOLOE (Real-time seeing anything) (Wang et al., 2025) has been developed for open-vocabulary segmentation and detection tasks. YOLOE can detect any object class in real time by utilizing prompts from text, images, or its internal vocabulary. Drawing from YOLOv10 and inspired by YOLO-World, YOLOE significantly enhances accuracy and efficiency compared to existing zero-shot models. It retains the traditional YOLO architecture, featuring a neck for multi-scale fusion, a convolutional backbone for feature extraction, and an anchor-free, decoupled detection head that independently predicts objectness, classes, and bounding boxes. This YOLO model offers three detection modes: text-prompted, visual-prompted, and prompt-free. The pre-trained models YOLOE-11S, YOLOE-11M, YOLOE-11L, YOLOE-v8S, YOLOE-v8M, and YOLOE-v8L are available along with their respective pre-trained weights for evaluation purposes.

Although there are several YOLOE variations with pre-trained weights, YOLOE-v8 was chosen for this investigation. The choice was based on its reduced model size and faster inference speed, while

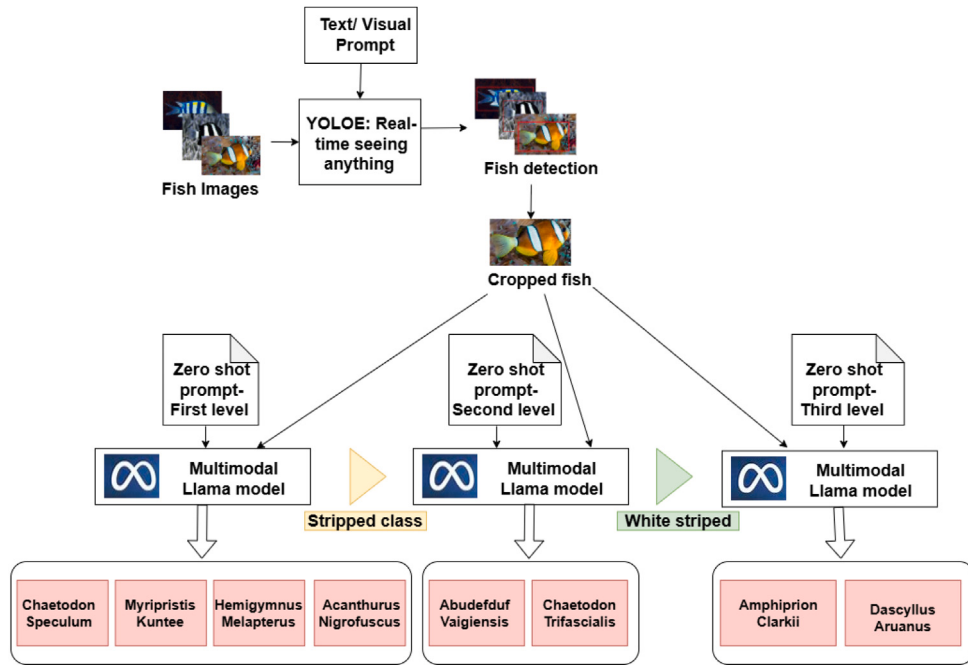


Fig. 1. Block diagram of the proposed system.

Table 1
Details of fish species used.

Species	Image	Number of images	Total fish count
<i>Abudefduf Vaigiensis</i>		17	19
<i>Acanthurus Nigrofusus</i>		11	11
<i>Amphiprion Clarkii</i>		14	17
<i>Chaetodon Speculum</i>		16	17
<i>Chaetodon Trifascialis</i>		16	16
<i>Dascyllus Aruanus</i>		15	17
<i>Hemigymnus Melapterus</i>		15	15
<i>Myripristis Kuntze</i>		10	13

still delivering satisfactory zero-shot detection performance for the required application. The YOLOE-v8s model was evaluated using both visual and text prompting for the analysis. While visual prompts offer reference images that assist the model in identifying visually comparable things without explicit textual input, text prompts use natural language descriptions of target objects to guide the model, enabling zero-shot identification based on semantic understanding. The prompt text “fish” was provided to the model for text prompting. On predicting fish on the image, each image is associated with a bounding box and a confidence score. The detection procedure was also evaluated using visual prompting. A single typical image was manually annotated with a bounding box for visual prompting in order to give the YOLOE model the visual reference it needs. The reference image and

its bounding boxes are passed along with the input images on which detection is to be performed. Based on the visual prompt, the YOLOE model detects objects in the given images. The eight-class fish species dataset is used to assess the effectiveness of both methods. Furthermore, the effectiveness of zero-shot learning with YOLOE was assessed on additional public datasets, like the medical pill dataset (Cho et al., 2025) and the African wildlife dataset (Ferreira, 2020).

2.4. Multimodal large language model

Multimodal Large Language Models (MLLMs) (Wu et al., 2023) work with text and other types of data, including images, audio, and video, to handle and produce content across various modalities. They combine the sophisticated reasoning skills of LLMs with the capacity to acquire, process, and produce multimodal information. An MLLM can accept diverse inputs like images, audio, video, code, and text, as shown in Fig. 2. Task-specific instructions direct the model on the necessary operation, such as classification, explanation, captioning, or cross-modal reasoning. The model generates a suitable response based on these inputs and the given instructions; depending on the task, the answer may be text, phonic, or code-based. This multimodal interaction capacity is crucial to the work as zero-shot understanding and categorization are made possible by the proposed system, which uses both visual inputs and textual prompts.

Llama4 MLLM (Meta, 2025) is utilized in the multilevel fish species classification process. The Llama4 Models are a group of pretrained and instruction-tuned mix-of-experts. This LLM comes in two sizes: Llama4 Scout and Llama4 Maverick. Coding, tool-calling, multimodal comprehension, multilingual activities, and agentic systems are all excellent uses for these models. Llama4 used a Mixture-of-Experts (MoE) design and native multimodality to differentiate itself from its predecessors. MoE, which has a collection of specialized mini-models known as experts, only activates a few relevant mini-models for each token. This architecture results in a total number of parameters in Llama4 Maverick to 400B while keeping the active parameters per inference at 17B per token.

In Llama4, a few of the transformer’s feed-forward levels are replaced by expert layers. For every input token, a gating network chooses which expert or experts to use. Every expert has its own

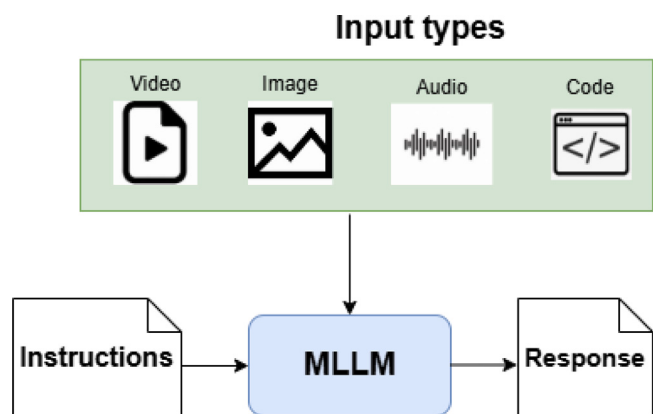


Fig. 2. Basic architecture of MLLMs.

weights and is a smaller feed-forward network. The gating function sends each token to the most relevant expert. The output of the MoE layer is the sum of the expert outputs for each token. Significant computational savings are achieved using this sparse activation, and the gating ensures that most experts stay dormant for a particular token.

MLLMs significantly contribute to zero-shot fish species categorization without the need for model training. The cropped fish image serves as the visual input for the model, while the textual input provides a structured prompt outlining the unique traits of each species. The MLLM determines the most likely species by comparing the visual characteristics of the fish with the textual descriptions with multimodal alignment and reasoning. The suggested method can classify species even in the absence of labeled training samples, due to its capacity to link visual characteristics with descriptive text. This is especially useful for marine species with small image datasets.

2.5. Zero shot fish species classification

The Llama4 model receives a prompt using the physical traits of each species of fish. A brief description of a fish's appearance is given, focusing mostly on the color and patterns on its body. The species name and description are sent as a zero-shot prompt to the LLM model. The best description for the input image is then determined by analyzing the prompts and the given image. Based on the matching description, LLM will return the species type in the response.

The LLM was first given a single zero-shot instruction to classify all eight species simultaneously, but the result was noticeably inferior. Subsequent investigation showed that the LLM model's responses were sensitive to order. In particular, LLMs have a tendency to favor the species stated last, giving more weight to information that appears later in the prompt. The classification's accuracy and fairness were impacted by this spatial bias. A multi-stage classification approach, in which the species are categorized hierarchically rather than all at once, was used to address this problem. Different prompts are used for each of the three phases of the classification procedure.

2.5.1. First level classification

At the first level of classification, the species are separated into five different classes. The species *Abudefduf Vaigiensis*, *Amphiprion Clarkii*, *Chaetodon Trifascialis*, and *Dascyllus Aruanus* are together referred to as the "Stripped class" due to their stripped body patterns. Level 1 considers the following classes:

Acanthurus Nigrofusculus	Chaetodon Speculum	Hemigymnus Melapterus	Myripristis Kuntze	Stripped class
--------------------------	--------------------	-----------------------	--------------------	----------------

The descriptive prompt given to the LLM model for the identification of species at the first level is given below.

```

<description>
The species name and the appearance of the
fishes are given below. Fish species
descriptions:

1. Chaetodon Speculum- The fish have
yellow body with BIG BLACK ROUND
spot.
2. Stripped class- The fish have BLACK
AND WHITE lines on the body.
3. Acanthurus Nigrofusculus - The fish
have brown body with small colored
spots all over the head region.
4. Myripristis Kuntze- The fish have
orange body color with dark red
patch near gills.
5. Hemigymnus Melapterus- Front part of
the fish body is in white color
and lateral part is in black color
with yellow tail.

<instruction>
Analyze the given cropped image of a fish
and identify which species best matches
the given image based on the
appearance of the fish.

<response format>
Answer only after checking all the species
types mentioned. Return only the
species name. No need of explanation.
  
```

In order to prevent the LLM model from providing additional information, LLM was specifically instructed in the prompt to merely return the species name. Even so, LLM has occasionally given a detailed response, as seen below:

Step 1: Analyze the given image description The image shows a fish with a yellow body and a large black spot.

Step 2: Compare the image description with the given fish species description.

Step 3: Identify the matching species.

Based on the description of having a yellow body with a round black spot, Chaetodon speculum is the matching species.

The final answer is: Chaetodon speculum

Post-processing was hence conducted on the response provided by the LLM to get only the final response species name. where the species name is obtained from the full response provided by the LLM using a fuzzy-based text matching. If the fish image is categorized as being in the "Stripped class" after the first classification stage, a second level of classification is performed.

2.5.2. Second level classification

In the second stage of categorization, images are divided into three classes. When evaluating the categorization of the four species in the second level, the similarity between *Dascyllus Aruanus* and *Amphiprion Clarkii* was generating a significant amount of false positives. As a result, the second level was further separated into three groups, including a mixed class of *Dascyllus Aruanus* and *Amphiprion Clarkii* known as "White striped". In the second level, images are categorized into the following categories:

Abudefduf Vaigiensis	Chaetodon Trifascialis	White striped
----------------------	------------------------	---------------

The textual prompt given to LLM with the fish description for level 2 classification is given below.


```
<description>
The species name and the appearance of a
set of fishes are given below. Fish
Species Descriptions:

1. White striped- The fish have TWO
  PROMINENT white vertical stripes on
  the body.
2. Abudehdud vaigiensis- The fish have
  body with black and white stripes
  and yellow color near dorsal fin.
3. Chaetodon trifascial - The fish have
  body with small black and white
  stripes with yellow fins.

<instruction>
Analyze the given cropped image of a fish
and identify which species best matches
the given image based on the
appearance of the fish.

<response format>
Answer only after checking all the species
types mentioned. Return only the
species name. No need of explanation.
```

White striped class is further passed onto the third level of categorization, for the identification of the species type.

2.5.3. Third level classification

In the final level of classification, white striped class is categorized into the two classes:

Dascyllus aruanus	Amphiprion Clarkii
-------------------	--------------------

The two species’ distinctive white stripes made it difficult to tell them apart on the second level. As a result, a more detailed description of these fish’s physical characteristics is included in the classification prompt. The following is the zero shot prompt:

```
<description>
The species name and the appearance of a
set of fishes are given below. Fish
Species Descriptions:

1. Dascyllus aruanus - The fish have
  black and white striped body with a
  black fin. Have the stripes like
  that of a zebra.
2. Amphiprion Clarkii - The fish have
  TWO prominent white vertical
  stripes on the body with YELLOW
  colored head and tail.

<instruction>
Analyze the given cropped image of a fish
and identify which species best matches
the given image based on the
appearance of the fish.

<response format>
Answer only after checking all the species
types mentioned. Return only the
species name. No need of explanation.
```

A user interface (UI) is developed to enhance the usability of the categorization system. The UI was created using the Streamlit package (Khorasani et al., 2022), which accepts the user’s original fish image and detects fish using the YOLOE model. A reference template image

Fish Species Classification Agent

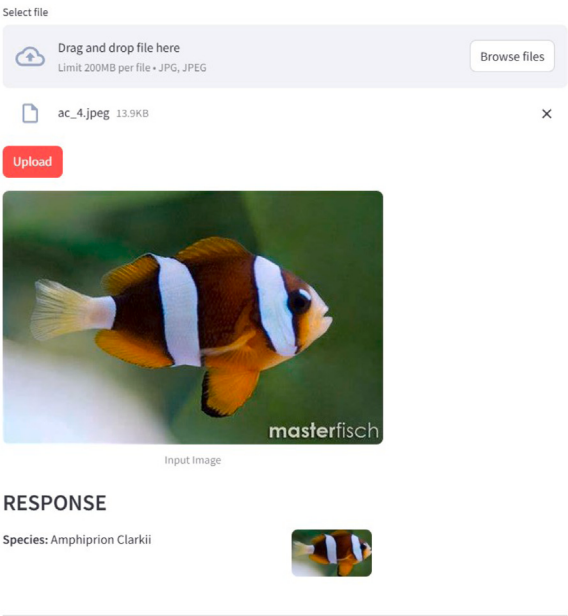


Fig. 3. UI screen of fish species classification application.

and the corresponding bounding boxes are provided to the YOLOE model. The user gets an error message if YOLO is unable to detect. If fish is detected by YOLO model, it is cropped and advanced to the first classification level when it is recognized by the YOLOE model. Depending on the class to which it is given, the image will also be sent to the second and third levels of categorization if necessary. The process is ultimately made easier by providing the user with a tabular response on the user interface screen that displays both the cropped fish image and the predicted species class. A screenshot of the user interface screen is shown in Fig. 3.

2.6. Evaluation parameters

Object detection combines image recognition with object localization to identify the presence of objects and draw bounding boxes around them. Conversely, classification is a supervised machine learning technique that predicts a class label based on given data. Precision, recall, F1-score, and accuracy are commonly used to assess object detection and classification techniques (Yao and Zhou, 2008). An additional metric used for evaluating object detection is the mean average precision (mAP) (Bashir and Porikli, 2006).

Precision is defined as the ratio of correctly detected objects, or actual positive detections, to all positive predictions, including false positives. Precision shows how well the model can distinguish relevant instances from all the instances. The ability of a model to identify every relevant object in a dataset is measured by recall. It is calculated by dividing the number of real objects by the number of true positive detections, which includes both true positives and false negatives. Recall measures how well the model locates every relevant object in the dataset. The F1-score is highly helpful when both false positives and false negatives are significant and a balance between precision and recall is required. The F1-score is the harmonic mean of precision and recall, with equal weight given to each statistic.

The most basic assessment metric is accuracy. It is computed as the proportion of accurate predictions to all predictions. mAP is a metric for evaluating the performance of object detection models. It calculates the mean for all classes after averaging the precision across a range of recall values for each class.

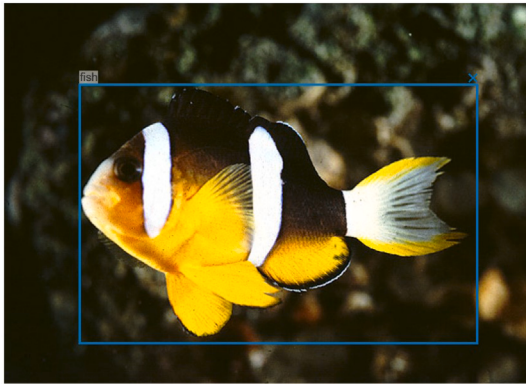


Fig. 4. Template image for visual prompting.

Table 2

Performance comparison of textual and visual prompting in YOLOE with different datasets.

Dataset	Prompt	Accuracy	Precision	Recall	F1-score	mAP
African wildlife	Text	56.51	56.51	73.07	63.73	59.22
	Visual	81.65	81.65	69.14	74.87	62.39
Medical pill	Text	92.56	92.56	17.62	29.61	18.13
	Visual	64.95	64.95	89.53	75.28	78.44
Fish dataset	Text	37.35	37.35	73.81	49.60	50.34
	Visual	88.00	88.00	69.84	69.84	59.69

3. Results and discussion

An 11th Gen Intel(R) Core(TM) i3 system without GPU was used to design the system. YOLOE was used using an IoU threshold of 0.45, a confidence threshold of 0.25, and an input image size of 640×640 . The YOLOE-v8s model was used for zero-shot inference without any further fine-tuning. The first analysis was carried out using the word “fish” as the text prompt. Later, analysis of YOLOE was also done using visual prompts. The YOLOE object detection model locates the fish region by processing the images. The sample template image used for visual prompting for fish identification is shown in Fig. 4. Other datasets, such as the African wildlife dataset and the medicinal pill dataset, are also used to evaluate the YOLOE model prompting methodologies. The African wildlife dataset comprises images of animals such as buffalo, elephants, rhinos, and zebras, whereas the medical pill dataset contains images of pharmaceutical tablets with annotations for applications including pharmaceutical quality assurance, pill sorting, and regulatory compliance. The performance evaluation of YOLOE with text and visual prompts for different datasets is shown in Table 2. YOLOE’s evaluation results on these datasets with both textual and visual prompting show that visual prompting can enhance detection.

After fish are detected, the fish region is clipped and submitted to the classification stage, where a zero-shot prompt MLLM model is employed. The study was conducted using GroqAPI’s models and the Huggingface (Jain, 2022) inference client. With a maximum token count of 100 and a temperature of 0.6, the LLM-based classification employed the Llama4 Maverick model. No training is done because the system is zero-shot. The average value, obtained from various temperature runs and slight prompt fluctuations, is what is reported. As explained in Section 2.5, the system employs three classification steps. Initially, all eight classes were classified simultaneously using a single zero-shot prompt with a description of each species. The performance was worse, with an F1 score of 56%, a recall of 64%, an accuracy of 62%, and a precision of 59%. Fig. 5 displays the classification confusion matrix for all eight species together. Upon analysis, it became clear that the LLM response had an order sensitivity problem. When LLMs make decisions, they tend to give more weight to recent context. The

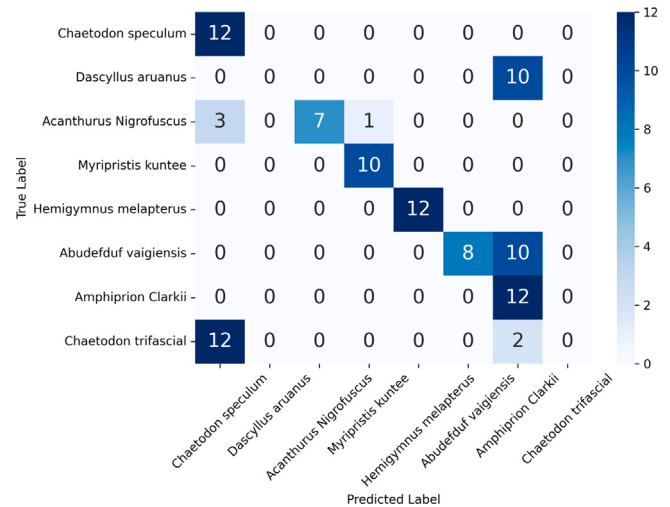


Fig. 5. Confusion matrix of all eight species classification in a single level.

comparison gives more weight to the species that was stated later in the prompt, which causes bias in classification. Therefore, the concept of dividing the classification process into several steps was taken into consideration.

In the first level of classification, cropped fish images are compared with descriptions of five classes. Four fish species with stripes on their bodies are grouped into a single class entitled *Stripped class*. The classification accuracy increased to 76% with 81% precision, 87% recall, and 80% F1-score when fish species with striped bodies were grouped into a single class. The confusion matrix for the first level classification step is shown in Fig. 6. The fish that are categorized as belonging to the *Stripped class* will proceed to the subsequent classification phase. It was discovered that it was difficult to differentiate between *Dascyllus aruanus* and *Amphiprion clarkii* when the four classes of the *Stripped class* were classified into separate classes simultaneously. These two fish species were therefore combined into a single class for the second level of classification, and an analysis was carried out. The classification stage’s performance has improved by combining *Dascyllus aruanus* and *Amphiprion clarkii* into a single class called *White striped*. Fig. 7 compares performance metrics for second-level classification with four and three classes. The confusion matrices of the two classification techniques assessed at the second level are displayed in Fig. 8. The results indicate that when *Dascyllus aruanus* and *Amphiprion clarkii* were grouped into a single class during classification, higher performance was obtained. For this reason, the second level of the proposed system used the three-class categorization.

Lastly, the *White striped* group of fish is categorized into species level at the third level of classification. The accuracy, precision, recall, and F1-score for the third-level classification were 91%, 93%, and 90%, respectively. The confusion matrix of the third-level classification stage is displayed in Fig. 9. We investigated the system across several temperatures and prompt variations in order to evaluate the stability of the suggested zero-shot LLM classifier with regard to randomness and prompt wording. Various temperature settings were employed, including 0.0, 0.3, 0.6, and 1.0. Few minor changes in the prompts were also used for evaluation. The evaluation metrics mean and standard deviation were calculated after several runs with these modifications. To ensure that every class, regardless of class size, contributes equally to the final score, metrics were computed using macro-averaging (Gowda et al., 2021). Performance of the proposed classification pipeline is also evaluated using different MLLMs. Table 3 makes it clear that the Llama4 Maverick model performed better than the other MLLM models. Due to the restricted model accessibility in GroqAPI and HuggingFace,

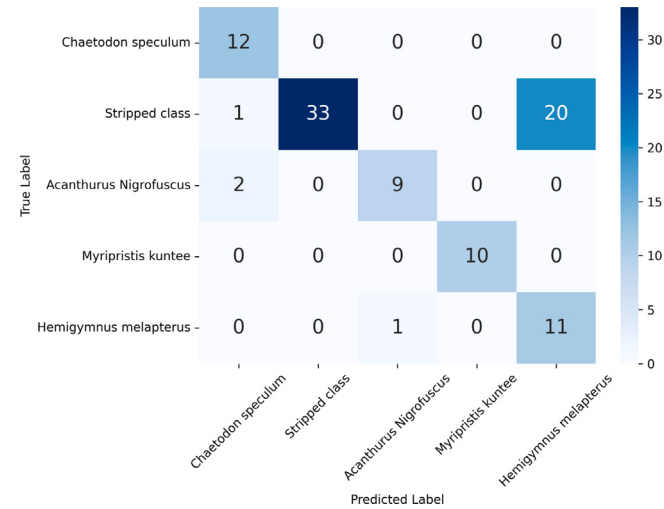


Fig. 6. Confusion matrix of first level classification of fish species.

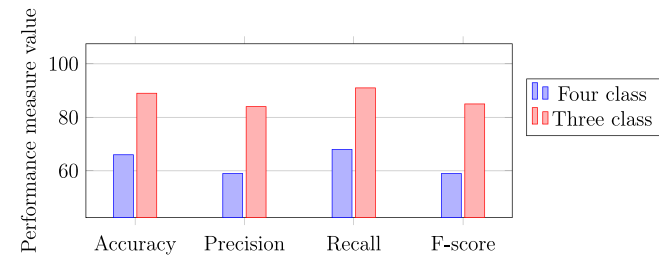


Fig. 7. Comparison of performance measures of four class and three class categorization in level2.

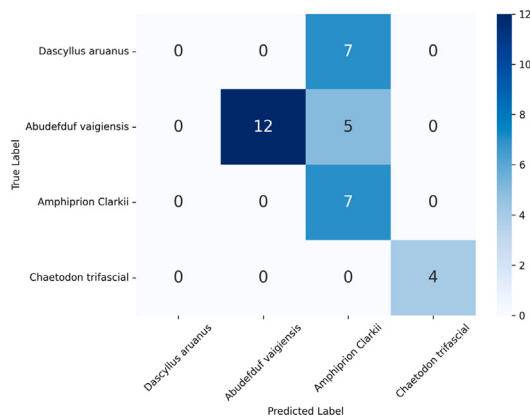


Fig. 8. Confusion matrix of second level classification into (a) Four class (b) Three class.

the evaluation was limited to these models. The confusion matrix of the top-performing MLLM model in the multistage hierarchical classification is shown in Fig. 10. The study demonstrates that zero-shot categorization using MLLM models can yield comparable outcomes. The suggested method has the potential to be applied to the categorization of fish species with low image availability because it functions in a zero-shot environment without requiring training data. For uncommon species with limited training datasets, this can be especially helpful.

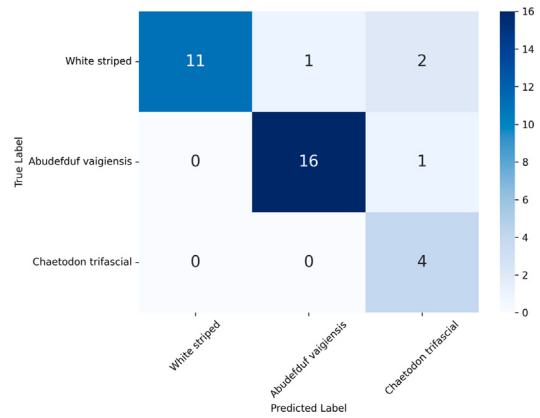


Fig. 8. (continued).

Table 3

Comparison of overall performance of the proposed system using different MLLMs.

MLLM model	Accuracy	Precision	Recall	F1-score
Llama4 Maverick 17B	87.20	89.4	87.4	86.8
Llama4 Scout 17B	68	76	68	67
Gemma3 27B	78	85	80	75
Qwen2.5VL 32B	49	48	45	40
Mistral Small3.1 24B	31	47	34	33

3.1. Comparison with state-of-the-art-systems

The same dataset was used to assess three baseline models in order to benchmark the proposed framework. Table 4 summarizes the performance of:

1. CLIP ViT-L/14 (Dong et al., 2022) in a zero-shot setting,
2. EfficientNet-B0 (Yang et al., 2021) trained with data augmentation
3. ResNet-50 (Koonce, 2021) trained with data augmentation.

For the evaluation, the CLIP model was employed with the same species description prompt as the proposed approach. Only 54% accuracy was attained by the CLIP zero-shot model, suggesting that fine-grained fish species detection requires more than just zero-shot image-text embedding alignment. The k-fold cross-validation method (Anguita et al., 2012) was applied with k of 5 to assess supervised learning systems. Augmentations like horizontal flipping, rotation, and brightness adjustments were applied on the original images for creation of augmented images for training process. EfficientNet-B0 was trained using the Adam optimizer (Tato and Nkambou, 2018), batch size 16, learning rate of 0.001, and 20 epochs. The macro-average accuracy was 84.31% with a standard deviation of 0.07. ResNet-50, on the other hand, obtained 89.55% accuracy with a standard deviation of 0.0214 after 20 training epochs using the Adam optimizer, a batch size of 16, and a learning rate of 0.001.

With no labeled training data, the proposed zero-shot pipeline outperformed the CLIP zero-shot baseline and was close to fully supervised performance with an accuracy of $87.20 \pm 0.013\%$. This demonstrates the effectiveness of multimodal reasoning in MLLMs for fine-grained recognition without the need for annotation or retraining.

3.2. Limitations and future scope

Although the proposed zero-shot fish classification system yields promising results, certain limitations must be recognized. First, changes in underwater imaging parameters, including turbidity, illumination, camera angle, and background complexity, might impact the model's

Table 4
Comparison of the proposed system with state-of-the-art-systems.

Model	Type	Training data required	Works on new species	Accuracy
CLIP ViT-L/14	Vision-Language	No	Immediate extension	54
EfficientNet-B2	CNN	Yes	Need retraining	84.31
ResNet-50	CNN	Yes	Need retraining	89.55
Proposed system	Multimodal LLM	No	Immediate extension	87.20

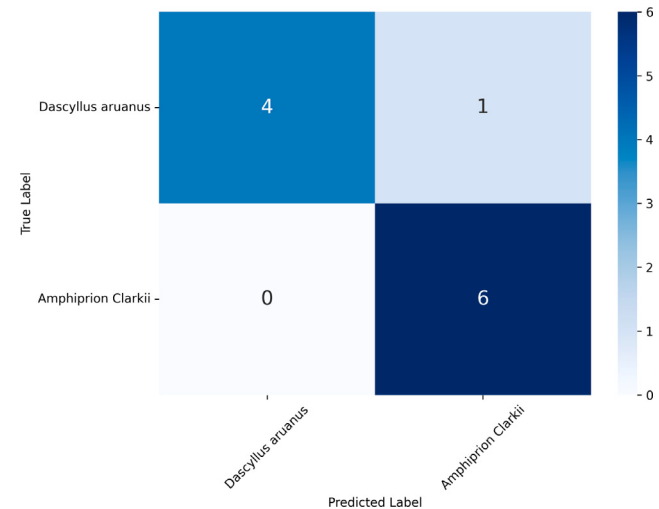


Fig. 9. Confusion matrix of third level classification of fish species.

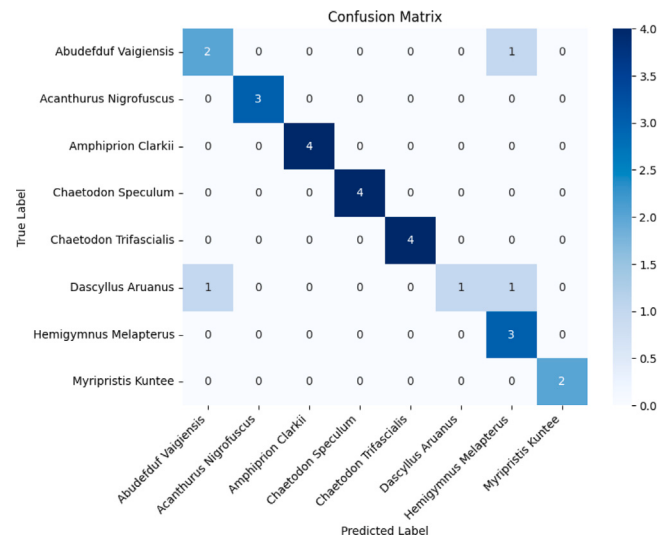


Fig. 10. Confusion matrix of overall prediction of fish species into eight classes using Llama4 Maverick model.

capacity for generalization. Second, the prompt-sensitive nature of zero-shot LLM-based classification necessitates careful prompt engineering, as descriptive variations might affect model predictions. Third, sequential detection and LLM reasoning are part of the inference pipeline, which could lengthen processing times and make it less appropriate for real-time applications in field deployments. The current system achieves an average latency of 2.5 s per image, which may pose a challenge for real-time underwater monitoring applications. Future work will focus on improving robustness across diverse underwater environments, reducing prompt dependency, and optimizing inference latency for real-time deployment on edge devices.

4. Conclusion

Manual fish species identification is typically labor-intensive and requires expertise. The advancement of computer vision technology has made the automatic separation of fish species simple. Since a large training dataset is necessary for the creation of such a classification system, it is difficult to classify species with little data. To get around this limited or no database problem, the proposed method suggests a zero shot fish species classification approach. A YOLOE model that uses visual prompt is first used to detect fish from the input images. Following detection, fish regions are cropped before being fed into a multi-level classification algorithm.

Zero-shot prompting with multimodal large language models is used in the suggested multistage hierarchical categorization method. Each input fish image is initially categorized into one of five main groups by the pipeline. To mitigate early-stage confusion, four visually similar species are combined into a single grouped class at the first level of the hierarchy. Images that belong to this grouped class are sent to the second level, where the grouped set is divided into three subclasses. Two species are united into a single “white-striped” subclass while the remaining species make up the other two subclasses. The white-striped subclass is further divided into the specific species in the last stage. By gradually managing visual ambiguity, this hierarchical grouping technique lowers cascade errors and improves total classification accuracy.

Accuracy, precision, recall, and F1-score were analyzed for each classification level. Classes are categorized into first and second levels based on the analysis. Five distinct large language models are used to evaluate the entire process, and the results are compared. The results show that Llama4 Maverick models topped other models in the multi-stage hierarchical classification approach. On the dataset employed, the accuracy of the proposed system was 87.2%; however, the performance of the LLM is strongly influenced by the fish description provided.

Beyond experimental findings, the suggested zero-shot methodology has great promise for practical uses such automated fish stock assessment, aquaculture surveillance, and marine biodiversity monitoring. These conclusions, however, are predicated on a particular dataset. Therefore, additional field-level validation across various underwater conditions, camera configurations, and species distributions is needed to guarantee dependable deployment. The descriptive prompt used in the study mainly depends on physical appearance, especially colors and patterns, which could result in predictions that are not correct in low light. Future research will examine more structurally described prompts that are better and can get around the lighting problem.

CCRediT authorship contribution statement

Jisha Anu Jose: Writing – original draft, Software, Methodology, Investigation. **Renjith Thomas:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., et al., 2012. The K' in K-fold cross validation. In: *Esann*. Vol. 102, pp. 441–446.
- Bashir, F., Porikli, F., 2006. Performance evaluation of object detection and tracking systems. In: *Proceedings 9th IEEE International Workshop on PETS*. pp. 7–14.
- Cadrin, S.X., Friedland, K.D., 1999. The utility of image processing techniques for morphometric analysis and stock identification. *Fish. Res.* 43 (1–3), 129–139.
- Cho, D.S., et al., 2025. Medical-pills detection using YOLOv11: A proof-of-concept study for pharmaceutical automation. Preprint.
- Christensen, J.H., Mogensen, L.V., Galeazzi, R., Andersen, J.C., 2018. Detection, localization and classification of fish and fish species in poor conditions using convolutional neural networks. In: *2018 IEEE/OES Autonomous Underwater Vehicle Workshop*. AUV, IEEE, pp. 1–6.
- Dong, X., Bao, J., Zhang, T., Chen, D., Gu, S., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., 2022. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*.
- dos Santos, A.A., Gonçalves, W.N., 2019. Improving Pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecol. Inform.* 53, 100977.
- Ferreira, B., 2020. African wildlife. <https://www.kaggle.com/biancaferreira/african-wildlife>. (Accessed 26 May 2025).
- Froese, R., Pauly, D., et al., 2010. FishBase.
- Ghaly, A., Ramakrishnan, V., Brooks, M., Budge, S., Dave, D., 2013. Fish processing wastes as a potential source of proteins, amino acids and oils: A critical review. *J. Microb. Biochem. Technol.* 5 (4), 107–129.
- Gong, B., Dai, K., Shao, J., Jing, L., Chen, Y., 2023. Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer. *Heliyon* 9 (6).
- Gowda, T., You, W., Lignos, C., May, J., 2021. Macro-average: rare types are important too. *arXiv preprint arXiv:2104.05700*.
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: A review. *Neurocomputing* 187, 27–48.
- Hridayami, P., Putra, I.K.G.D., Wibawa, K.S., 2019. Fish species recognition using VGG16 deep convolutional neural network. *J. Comput. Sci. Eng.* 13 (3), 124–130.
- Jain, S.M., 2022. Hugging face. In: *Introduction To Transformers for NLP: With the Hugging Face Library and Models To Solve Problems*. Springer, pp. 51–67.
- Jose, J.A., Kumar, C.S., 2020. Genus and species-level classification of wrasse fishes using multidomain features and extreme learning machine classifier. *Int. J. Pattern Recognit. Artif. Intell.* 34 (11), 2050028.
- Jose, J.A., Kumar, C.S., Sureshkumar, S., 2022. Region-based split Octonion networks with channel attention module for tuna classification. *Int. J. Pattern Recognit. Artif. Intell.* 36 (07), 2250030.
- Kandimalla, V., Richard, M., Smith, F., Quirion, J., Torgo, L., Whidden, C., 2022. Automated detection, classification and counting of fish in fish passages with deep learning. *Front. Mar. Sci.* 8, 823173.
- Khorasani, M., Abdou, M., Fernández, J.H., 2022. Web application development with streamlit. *Softw. Dev.* 498–507.
- Knausgård, K.M., Wiklund, A., Sjørdalen, T.K., Halvorsen, K.T., Kleiven, A.R., Jiao, L., Goodwin, M., 2022. Temperate fish detection and classification: a deep learning based approach. *Appl. Intell.* 52 (6), 6988–7001.
- Koonce, B., 2021. ResNet 50. In: *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*. Springer, pp. 63–72.
- Malla, S., Meena, M., Reddy, O., Mahalakshmi, R., Balobaid, A., 2022. A study on fish classification techniques using convolutional neural networks on highly challenged underwater images. *Int. J. Recent Innov. Trends Comput. Commun.* 10 (4), 01–9.
- Meta, A., The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, checked on 4 (7) 2025. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- Pratt, S., Covert, I., Liu, R., Farhadi, A., 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15691–15701.
- Rachmatullah, M.N., Supriana, I., 2018. Low resolution image fish classification using convolutional neural network. In: *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications*. ICAICTA, IEEE, pp. 78–83.
- Si, G., Xiao, Y., Wei, B., Bullock, L.B., Wang, Y., Wang, X., 2023. Token-selective vision transformer for fine-grained image recognition of marine organisms. *Front. Mar. Sci.* 10, 1174347.
- Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P., 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* 55 (13s), 1–40.
- Tamou, A.B., Benzinou, A., Nasreddine, K., Ballihi, L., 2018. Underwater live fish recognition by deep learning. In: *Image and Signal Processing: 8th International Conference, ICISP 2018, Cherbourg, France, July 2–4, 2018, Proceedings* 8. Springer, pp. 275–283.
- Tato, A., Nkambou, R., 2018. Improving adam optimizer.
- Tejaswini, H., Pai, M.M., Pai, R.M., 2024. Automatic estuarine fish species classification system based on deep learning techniques. *IEEE Access*.
- Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G., 2025. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*.
- Wang, M., Liu, M., Zhang, F., Lei, G., Guo, J., Wang, L., 2018. Fast classification and detection of fish images with YOLOv2. In: *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans. OTO*, IEEE, pp. 1–4.
- Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S., 2023. Multimodal large language models: A survey. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 2247–2256.
- Wu, J., Tang, X., Yang, Z., Hao, K., Lai, L., Liu, Y., 2025. An experimental evaluation of LLM on image classification. In: *Australasian Database Conference*. Springer, pp. 506–518.
- Yang, S., Shang, Z., Wang, Y., Deng, D., Chen, H., Cheng, Q., Wu, X., 2024. Data-free multi-label image recognition via LLM-powered prompt tuning. *arXiv preprint arXiv:2403.01209*.
- Yang, L., Yu, H., Cheng, Y., Mei, S., Duan, Y., Li, D., Chen, Y., 2021. A dual attention network based on efficientNet-B2 for short-term fish school feeding behavior analysis in aquaculture. *Comput. Electron. Agric.* 187, 106316.
- Yao, Y., Zhou, B., 2008. Micro and macro evaluation of classification rules. In: *2008 7th IEEE International Conference on Cognitive Informatics*. IEEE, pp. 441–448.
- Zhu, J., Yin, S., Liu, X., Wang, X., Yang, Y.-H., 2025. FishDetectLLM: Multimodal instruction tuning with large language models for fish detection. *Knowl.-Based Syst.* 318, 113418.